Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Subject Section

GeneLab: Omics database for spaceflight experiments

Shayoni Ray[1*], Samrawit Gebre[2*], Homer Fogle[2], Daniel C. Berrios[1], Peter B. Tran[3], Jonathan M. Galazka[4], Sylvain V. Costes[4**]

*Both the authors contributed equally
[1] Space Biosciences Division, USRA/NASA Ames Research Center, Moffett Field, CA 94035, USA
[2] Space Biosciences Division, KBRwyle/NASA Ames Research Center, Moffett Field CA 94035, USA
[3] Intelligent Systems Division, NASA Ames Research Center, Moffett Field, CA 94035, USA
[4] Space Biosciences Division, NASA Ames Research Center, Moffett Field, CA 94035, USA

** To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Running head: GeneLab: The spaceflight omics database

Abstract

Motivation – To curate and organize expensive spaceflight experiments conducted aboard space stations and maximize the scientific return of investment, while democratizing access to vast amounts of spaceflight related omics data generated from several model organisms.
Results - The GeneLab Data System (GLDS) is an open access database containing fully coordinated and curated "omics" (genomics, transcriptomics, proteomics, metabolomics) data, detailed metadata and radiation dosimetry for a variety of model organisms. GLDS is supported by an integrated data system allowing federated search across several public bioinformatics repositories. Archived datasets can be queried using full-text search (e.g., keywords, Boolean and wildcards) and results can be sorted in multifactorial manner using assistive filters. GLDS also provides a collaborative platform built on GenomeSpace for sharing files and analyses with collaborators. It currently houses 172 datasets and supports standard guidelines for submission of datasets, MIAME (for microarray), ENCODE Consortium Guidelines (for RNA-seq) and MIAPE Guidelines (for proteomics).
Availability and Implementation - https://genelab.nasa.gov/
Contact - sylvain.v.costes@nasa.gov

1 Introduction

Six decades of research on the effects of long- and short-term missions to space have helped isolate and quantify the impact of extraterrestrial factors on biological systems. These factors include exposure to microgravity, increased levels of ionizing radiation, elevated concentrations of carbon dioxide and psychological stressors on biological systems. Among these factors and systems, effects of microgravity and radiation on rodents, microbes and plants are the best studied. The critical effects of microgravity studied in rodents include defects in hepatic lipid metabolism (Jonscher, et al., 2016), bone loss due to osteoclastic degradation and osteocytic osteolysis (Blaber, et al., 2013), alterations in myogenic cell proliferation and differentiation, calcium homeostasis and muscle development (Allen, et al., 2009; Gambara, et al., 2017) along with defects in immune response including lymphocytopenia, compromised response of lymphocytes to stimulating agents and increased aberrations in lymphocyte DNA (Pecaut, et al., 2017; Worthington, et al., 2012). Previous investigations using bacteria such as the enteric pathogen *Salmonella typhimurium* demonstrated that simulated microgravity can induce increased resistance to environmental stresses (acid, osmotic, and thermal), increased survival in macrophages and increased virulence in a murine model (Wilson, et al., 2007). Microgravity has also been found to impede root hair development in *Arabidopsis thaliana* and thus negatively affecting nutrient absorption (Kwon, et al., 2015). The effects of space ionizing radiation, consisting of high charge and energy (HZE) particles (also referred as galactic cosmic rays - GCR), on astronauts is a critical concern for travel beyond the Earth's magnetosphere (Koike, et al., 2005). The high linear energy transfer

(LET) characterizing most GCR particles lead to enhanced deleterious effects in tissues, including progressive vascular changes (Weintraub, et al., 2010), cognition and behavioral effects (Rabin, et al., 2009), higher cancer incidence and visual impairment. For example, cognitive function characterized by fatigue and decline in performance have been observed in rodents exposed to simulated space radiation (Chancellor, et al., 2014). It is believed that DNA damage is one of the root cause of these effects leading to genetic mutation, which is considered to be an initiating event for carcinogenesis (Heuskin, et al., 2016; Sridharan, et al., 2015; Tang, et al., 2015). However other damaged cellular substructures such as mitochondria (Hei, 2014) or centrosome (Maxwell, et al., 2008) have also been shown to contribute to the chronic induction of oxidative stress and genomic instability following low doses of HZE. Additionally, damages to lipids and proteins may play an important role in the disruption of the microenvironment (Andarawewa, et al., 2011; Barcellos-Hoff and Ravani, 2000), thus accelerating pre-existing conditions associated with aging (e.g. cancer, cardiovascular disease, cataract formation). In the case of cancer, radiation affecting targets other than DNA are referred to as non-targeted effects (NTE) and are believed to transform the surrounding microenvironment, leading to cancer promotion (Barcellos-Hoff and Ravani). For example, space radiation can elicit chronic inflammation, which in turn disrupt tissue homeostasis affecting cellular interactions and increasing oxidative-process-initiated carcinogenic mutation (Barcellos-Hoff, et al., 2015). In *Arabidopsis thaliana*, both gamma radiation and HZE particles were found to affect plant growth and development by producing radicals that resulted in differential response to DNA double strand breaks (DSBs) (Missirian, et al., 2014). So far, although no studies have been conducted to assess the synergistic effects of ionizing radiation, microgravity and stress response in a confined environment; such investigations using different model systems are warranted for a deeper understanding of the effects of spaceflight environment.

Compared to a typical ground experiment, which involves larger sample size and greater number of personnel performing the experiments, spaceflight research imposes major limitations on the amount of experimentation that can be conducted; the size of and access to experiment equipment and the availability of trained staff and supplies needed to perform the scientific work. Furthermore, there are significant technical challenges encountered when developing experiment hardware to operate in low gravity, optimizing retrieval, processing and storage of biological samples. Thus it is particularly important to maximize the scientific value of specimens recovered from each spaceflight experiment. The GeneLab project is part of an effort from NASA to optimize scientific return of investment (ROI) on any experimental payload and democratize the access to the vast amounts of omics (e.g. genomics, transcriptomics, proteomics, and metabolomics) data produced from such specimens. Since it was launched in 2015, GeneLab (https://genelab.nasa.gov) has been providing researchers across the globe unrestricted access to a wide array of fully coordinated and curated raw omics data (e.g. genomics, transcriptomics, proteomics, and metabolomics) and associated metadata from spaceflight missions and ground simulations of space environments. In addition to the omics database, the platform data systems feature federated search across other open omics repositories and collaborative data analysis tools.

Such a model of open access science repository enables NASA to fulfill the following goals:

(1) Develop an integrated repository and bioinformatics data system for analysis and in-silico modeling – GeneLab data system (GLDS) offers large amount of fully curated data, the computational tools and workspace to create a public and scientific community resource for predictive modeling for future hypothesis-driven space biology research.

(2) Enable the discovery and validation of molecular networks that are influenced by space conditions through ground-based and flight research using next-generation ‘omics' technologies – GLDS offers an organized platform for deposition, curation, analysis and visualization of complex multi-parametric space-flight data emanating from a host of biological model systems.

(3) Engage the broadest possible community of researchers, industry, and the general public to foster innovation – GeneLab strives to drive and sustain scientific collaborations from universities, industries, students, PIs as well as citizen scientists to further the understanding of the space effects on living systems and enable the highest diversity of scientific endeavors.

(4) Strengthen international partnerships by leveraging existing capabilities and data sharing – GeneLab offers the prospect of collaboration with several national and international space agencies, pharmaceutical industries, universities and engineering firms for development of space hardware and software to standardize processes of data collection and analysis and hence increase relevancy of the space-flight missions.

The GeneLab project and the repository thus helps scientific investigators around the globe, to synergize space-based research conducted at multiple levels of complexity on model organisms, resulting in creation of minable big data. These results can then be utilized to enhance the pace of high-impact scientific discoveries to advance technology and medical development for space exploration. In this article, we describe the structure and content of the GLDS and the on-going work in creating a centralized system for spaceflight-relevant omics data storage, analysis and visualization.

2 Systems and Methods

2.1 Database Content

GeneLab has emerged as an omics data archive that is accepted by the space biology community. Furthermore, recent NASA Research Announcements (NRAs) require deposition of data from proposals selected for funding into the GeneLab database. As of July 2018, GLDS houses 172 studies of several spaceflight factors such as gravity, electromagnetic fields (Mayer-Wagner, et al., 2018), atmospheric pressure, temperature and ionizing radiation, as well as confounders such as age, diet and exercise on several biological model systems (plants, in-vitro cell cultures, bacteria, yeast, fruit flies, worms, mice and rats and humans). GeneLab and its sister database, NASA's LSDA (Life Sciences Data Archive), are the only open-access databases focusing exclusively on space life sciences, with experimental data collected through numerous space research organizations around the world. Over half of the experiments housed in GeneLab were conducted in space, collected from experiments performed on Space Transport System (the Space Shuttles), the International Space Station, parabolic flights, and free-flying satellites (Figure 1). In addition to data from experiments conducted in space, GeneLab includes ground studies simulating spaceflight environments (primarily microgravity and cosmic radiation). Examples of microgravity simulation platforms include 2-D clinostats, rotating wall vessels, or random positioning machines (Herranz, et al., 2013). Ground-based radiation experiments provide knowledge of the effects of cosmic rays on astronaut health, increasingly important as human spaceflight expands beyond low-earth orbit (LEO). In recent months, GeneLab has increased its focus on including datasets with ionizing radiation as factors, ranging in total dose and exposure time (Beheshti, et al., 2018). In fact, 40% of GeneLab data is currently radiation data, either alone or in conjunction with microgravity, ground or space, with majority of experiments conducted with gamma exposure followed by proton,

^{56}Fe, ^{16}O, ^{12}C, ^{28}Si and neutrons (Beheshti, et al., 2018). For space experiments investigating the effect of microgravity or other stressors, GeneLab additionally has an extensive catalog of parameters specifying experimental duration, and the average, minimum and maximum absorbed dose received during the entire experiment, for samples aboard ISS, BION-M1 and FOTON satellites. This can be found under the 'Environmental Data' section on GLDS: https://genelab-data.ndc.nasa.gov/genelab/environmental/radiation

GeneLab first sought out "legacy" (i.e., pre-GeneLab) omics data from space biology investigations that could be imported from other databases such as NCBI's Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) database, or from the European Bioinformatics Institute's ArrayExpress (AE), Nucleotide Archive (ENA) and PRoteomics IDEntifications (PRIDE) databases. A third of the datasets in the GeneLab repository consists of rodent experiments, followed by microbes, human cell lines and plants (Figure 1A). About three quarts contain data from transcription profiling assays (Figure 1B), and most were performed using a microarray platform (Figure 1D). Rapid development and cost reduction of high-throughput sequencing (RNA-Seq, bisulfite DNA sequencing, etc.) have enabled such kinds of data to become more prevalent in the database. GeneLab also includes studies with protein expression profiling, metabolite expression profiling and metagenomic data from spaceflight and space-analog investigations.

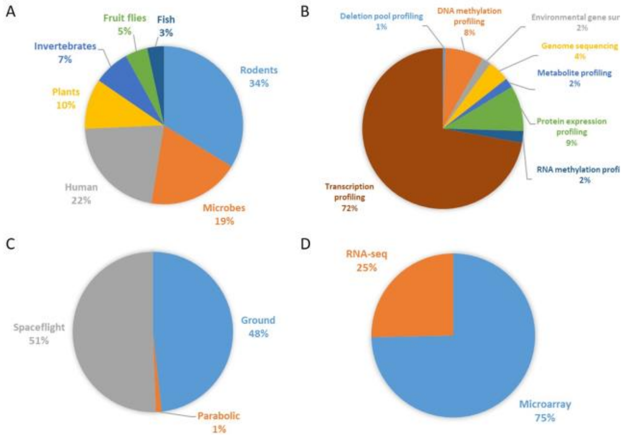


Figure 1. Database Content in GeneLab Repository. A graphical representation of the several types of spaceflight data currently curated in GeneLab database, by model organism (A), by assay type (B), by study type (C) and by transcriptomic assay type (D).

2.2 Database Features

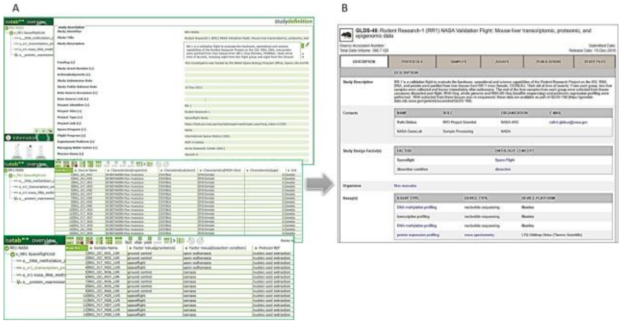
2.2.1 Metadata and Data files

With the increasing amount and complexity of omics data being generated, there is substantial value in the use of community-defined, common models of metadata and terminology so that omics data and results are discoverable and reliably reproducible. GeneLab uses the ISA-Tab specification (Sansone, et al., 2012) and semantic model (Neumann, 2005) for organizing and representing omics metadata. The ISA framework is built on a semantic model of the investigation, study, and assay metadata categories, and is extensible for domain-specific metadata. Using GeneLab's extensions to the ISA-Tab specification, the investigation category can provide programmatic details such as title, description, funding source, project type (spaceflight or ground study), space mission information, resulting publications, and contact information. The study category includes metadata for biosample characteristics and protocols, and per-sample experimental factor values. Finally, the assay category can include metadata for assay protocol parameters and the assay data files associated with each sample.

Investigators submitting data to GeneLab are encouraged to use the ISAcreator tool (Figure 2A), a java desktop application, bundled with the GeneLab-specific ISA-Tab extensions, to create, edit, and visualize ISA-Tab files (Sansone, et al., 2012). The ISAcreator tool has integrated ontology search and selection capabilities, so that users can use common, domain-specific, vocabulary terms when creating metadata. The attributes for each dataset is available for public access through the GLDS repository site (https://genelab-data.ndc.nasa.gov/genelab/projects?page=1&paginate_by=25). For each dataset, the metadata is viewable through a set of tabbed user interfaces, including the following tabs: Description, Protocols, Sample table, Assay table, Publications, and Study files (Figure 2B). In order to download metadata and/or data files, users need to navigate to the Study Files tab and click the desired file links. Each file is categorized by file types and assay types.

The following information can be found in each tab:

- *Description tab* provides an overview of the study including a text description, contact information, experimental factors, organism(s), type of assay, project and/or mission details and funding information.
- *Protocols tab* provides the names and descriptions of the sample collection, treatment, assay, data processing, and any other study protocols.
- *Samples and Assay(s) tabs* provide sample and assay level details formatted in a navigable table. Information found in these tabs include specific organism characteristics, study factors and treatments, radiation metadata, sample and sample processing metadata, and assay execution parameters.
- *Publications tab* contains metadata for relevant publication(s) including title, authors and link(s).
- *Study Files tab* provides metadata regarding raw and/or processed study data files. Each row includes information about the size, type and description of the files.



GLDS-48 Rodent Research-1 (RR1) NASA Validation Flight: Mouse liver transcriptomic, proteomic, and epigenomic data

Figure 2. Metadata Organization using ISAcreator. (A) GeneLab uses the ISAcreator tool to create, edit, and visualize ISA-Tab files allowing submitter to provide detailed information pertaining to description, experimental factors, organism(s) and type of assays, protocol, and project and/or mission details funding information as well as publications (B) Metadata for each dataset is viewable through a set of tabbed user interfaces, that includes the following tabs: Description, Protocols, Sample table, Assay table, Publications, and Study files

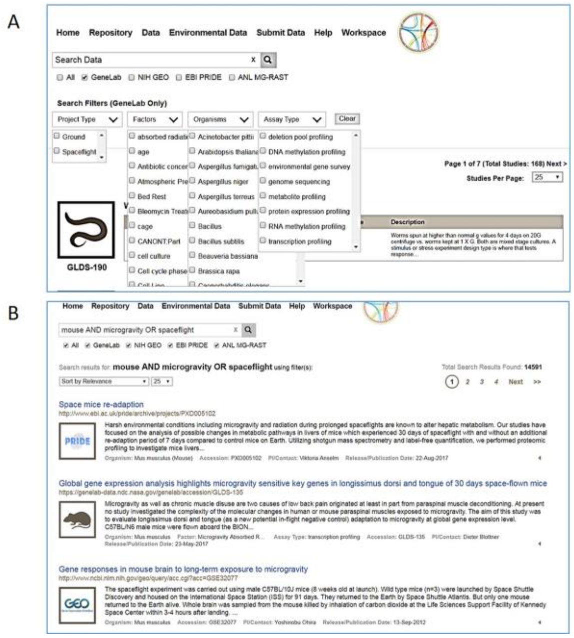
2.2.2 Searching

GeneLab Data System provides users a full-text search capability of its metadata. Full-text search terms can be a single word or multiple words with either Boolean, wildcards and/or string literals. Search results can be sorted by relevance, release date, source and title with either ascending or descending order. Filters can be applied to assist further in narrowing results by Project Type, Factors, Organisms, and Assay Types. The filter values for each category are dynamically pre-populated with metadata from the GeneLab-hosted data sets (Figure 3A).

Uniquely, GeneLab has federated with external databases rendering access to study level information, so users can search across multiple different platforms and omics using single search (Figure 3B). GeneLab is currently federated with:

- The National Institutes of Health (NIH) Gene Expression Omnibus (GEO)
- The European Bioinformatics Institute (EBI) Proteomics Identification (PRIDE)
- The Argonne National Laboratory (ANL) Metagenomics Rapid Annotations using Subsystems Technology (MG-RAST)

The GLDS does not duplicate the copies of the data sets found in the external databases; but instead routinely indexes the federated metadata attributes from the external data sets to keep the search content up-to-date.


A
Home Repository Data Environmental Data Submit Data Help Workspace
Search Data
□ All ☑ GeneLab □ NIH GEO □ EBI PRIDE □ ANL MG-RAST
Search Filters (GeneLab Only)
Project Type
Factors
Organisms
Assay Type
Clear
□ Ground
□ Spaceflight
□ absorbed radiati
□ age
□ Antibiotic concer
□ Atmospheric Pre
□ Bed Rest
□ Bleomycin Treat
□ cage
□ CANONT Part
□ cell culture
□ Cell cycle phase
□ Cell Line
□ Acinetobacter pittii
□ Arabidopsis thaliana
□ Aspergillus fumigat
□ Aspergillus niger
□ Aspergillus terreus
□ Aureobasidium pull
□ Bacillus
□ Bacillus subtilis
□ Beauveria bassiana
□ Brassica rapa
□ Caenorhabditis elegans
□ deletion pool profiling
□ DNA methylation profiling
□ environmental gene survey
□ genome sequencing
□ metabolite profiling
□ protein expression profiling
□ RNA methylation profiling
□ transcription profiling
Page 1 of 7 (Total Studies: 168) Next >
Studies Per Page: 25
GLDS-190
Description
B
Home Repository Data Environmental Data Submit Data Help Workspace
mouse AND microgravity OR spaceflight
☑ All ☑ GeneLab ☑ NIH GEO ☑ EBI PRIDE ☑ ANL MG-RAST
Search results for: mouse AND microgravity OR spaceflight using filter(s):
Total Search Results Found: 14591
Sort by Relevance
25
1 2 3 4 Next >>
PRIDE
GEO

Worms spun at higher than normal g values for 4 days on 20G centrifuge vs. worms kept at 1 X G. Both are mixed stage cultures. A stimulus or stress experiment design type is where that tests response...

Space mice re-adaption
http://www.ebi.ac.uk/pride/archive/projects/PXD005102
Harsh environmental conditions including microgravity and radiation during prolonged spaceflights are known to alter hepatic metabolism. Our studies have focused on the analysis of possible changes in metabolic pathways in livers of mice which experienced 30 days of spaceflight with and without an additional re-adaption period of 7 days compared to control mice on Earth. Utilizing shotgun mass spectrometry and label-free quantification, we performed proteomic profiling to investigate mice livers...
Organism: Mus musculus (Mouse) Accession: PXD005102 PI/Contact: Viktoria Anselm Release/Publication Date: 22-Aug-2017

Global gene expression analysis highlights microgravity sensitive key genes in longissimus dorsi and tongue of 30 days space-flown mice
https://genelab-data.ndc.nasa.gov/genelab/accession/GLDS-135
Microgravity as well as chronic muscle disuse are two causes of low back pain originated at least in part from paraspinal muscle deconditioning. At present no study investigated the complexity of the molecular changes in human or mouse paraspinal muscles exposed to microgravity. The aim of this study was to evaluate longissimus dorsi and tongue (as a new potential in-flight negative control) adaptation to microgravity at global gene expression level. C57BL/N6 male mice were flown aboard the BION...
Organism: Mus musculus Factor: Microgravity Absorbed R... Assay Type: transcription profiling Accession: GLDS-135 PI/Contact: Dieter Blottner Release/Publication Date: 23-May-2017

Gene responses in mouse brain to long-term exposure to microgravity
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE32077
The spaceflight experiment was carried out using male C57BL/10J mice (8 weeks old at launch). Wild type mice (n=3) were launched by Space Shuttle Discovery and housed on the International Space Station (ISS) for 91 days. They returned to the Earth by Space Shuttle Atlantis. But only one mouse returned to the Earth alive. Whole brain was sampled from the mouse killed by inhalation of carbon dioxide at the Life Sciences Support Facility of Kennedy Space Center within 3-4 hours after landing. ...
Organism: Mus musculus Accession: GSE32077 PI/Contact: Yoshinobu Ohira Release/Publication Date: 13-Sep-2012

Figure 3. Search Options in GeneLab Database. (A) When searching only in the GeneLab repository, users can use the drop-down filters to search for datasets. Users can select multiple terms to narrow down their search. (B) To search across multiple databases, users select the desired databases under the search bar and type in their search terms. The example shown above searches "All' databases using the search "mouse AND microgravity OR spaceflight". The search results include datasets from each databases and renders 14951 datasets.

2.2.3 Workspace

Leveraging from the GenomeSpace platform (Qu, et al., 2016), GeneLab has customized a collaborative workspace for file sharing and access to data analysis tools. Within this framework, users can share data files and analysis results with other users and/or groups with access controls, have access to the data in the GeneLab public repository, and data analysis tools. To access the workspace, first-time users will need to register an account.

Key capabilities:

- Share data and results with collaborators
- Import other publicly available data sources using convenient "Import from URL" feature
- Drag & drop files and documents from your desktop computer to workspace folders
- User and Group defined security access controls with Private, Shared, and Public folders and read, write, and delete file/folder operations
- View and navigate between GeneLab data listing and workspace environment
- Defaults to 30 GB quota of storage space per user. Additional storage space can be requested from the GeneLab team.

2.2.4 Submission

Process followed by data submitter:

Instructions for data submission can be found at https://genelab-data.ndc.nasa.gov/genelab/submissions. For detailed information on data and file formats, please refer to the table listed at: https://genelab.nasa.gov/help/faq/#5.

In brief, a submitter can create a GeneLab Workspace user account to transfer their metadata and data files. Then submitters use the ISAcreator tool, with the bundled GeneLab configuration file (see Step 3 - https://genelab-data.ndc.nasa.gov/genelab/submissions). This configuration file was created by GeneLab to denote all required metadata fields for spaceflight investigations, and is maintained by GeneLab with input from expert space biologists. The precise steps to follow for populating the required fields are found in the Submission Guide (https://genelab-data.ndc.nasa.gov/genelab/help/GeneLab_Submission_Guide_2.0.pdf)

Metadata should be populated using the ISAcreator tool and all applicable fields should be completed. The ISAcreator tool and tutorial, found in the link listed above will guide new submitters towards creating study description and experimental information for submission. Metadata, in ISA-tab format and raw data files, are uploaded to the submitter's private workspace folder and shared with the GeneLab Curation team for review. The GeneLab user manual contains more information on acceptable assay-specific metadata and compressed file formats to upload. Currently, we accept and publish open source file formats. Submitters are encouraged to convert raw and processed data files to common exchange formats whenever possible.

GeneLab follows the standard guidelines for submission and publication of datasets, such as MIAME (for microarray), ENCODE Consortium Guidelines (for RNA seq) and MIAPE Guidelines (for proteomics). File organization with metadata, raw and processed data files into directories, is critical to submission along with a desired release data. The terms and conditions for the GLDS repository can be found under: https://genelab-data.ndc.nasa.gov/genelab/terms/. The information uploaded on GeneLab database:

- Should include mention of product or service strictly in context of research
- Should not include classified, sensitive, proprietary or inappropriate information
- Should comply with export control regulations

After submission, principal investigators are provided with a unique GeneLab accession number. Additionally, GeneLab will begin issuing Digital Object Identifiers (DOI) for each dataset that will be provided to submitters and displayed to database users.

GeneLab curation process:

GeneLab meticulously curates (validates and maintains) each submitted spaceflight omics dataset. The process leverages the ISA-framework and adds specific fields for describing metadata for space-related omics experiments. GeneLab has also configured ISAcreator to include additional fields based on organism and assay type, which were suggested by the space biology community including omics experts who are a part of our Analysis Working Groups (AWG).

GeneLab curators review each dataset for completeness and accuracy of all field values. Curators also review the content of all associated publications for accuracy of experimental design descriptions, including materials and methods, sample information, assay information. When discrepancies between submitted and published metadata are discovered, we contact the Principal Investigator (PI) or submitters to verify. For each the above steps, curators seek guidance from the intramural GeneLab Science team composed of experts in space biology and bioinformatics.

All metadata are sent to the PI or data submitter for review. Once they confirm they are correct, a senior curator verifies and validates each dataset, and then releases the dataset for publication. Finally, each published data set goes through an internal scientific review by a panel of scientists, with expertise in spaceflight, omics, radiation, microgravity, etc. Consistency between expert reviews is maintained by discussion between the experts in assessing the scientific completeness of each dataset. The curation team provides this expert scientific feedback to the PI or data submitter, and, working with them, may further update the data set based on this feedback.

2.2.5 Contact

We have included detailed user guides on navigating the repository, data submission and general information about data types, technologies and other submission-specific queries. In case there are other questions regarding outreach, general FAQs, the GeneLab team can be reached via email distribution list at: genelab-outreach@lists.nasa.gov. For any publication that utilized data from GeneLab repository, we strongly encourage researchers to kindly acknowledge NASA GeneLab by citing this manuscript and by sending us reprints of manuscripts or oral presentations to help us assess our role in furthering the space biology research community.

3 Algorithm and Implementation

The GLDS is built using predominantly open source software components and tools. The web user interfaces are built using JavaScript frameworks, such as jQuery and AngularJS, runs on the Apache Tomcat Java-based application server, and search function is developed using ElasticSearch full-text search engine (Figure 4). The ISA-Tab formatted metadata is parsed into a standardized JavaScript Object Notation (JSON) format and stored in an object-oriented, NoSQL database built on MongoDB. GeneLab utilizes the NoSQL database due to its flexibility and scalability. To keep up with its growing data volume, the GLDS uses the Amazon Web Services (AWS) cloud infrastructure to host all of its data files and for any computing needs. The advantages of using AWS include scalability, flexibility in adding ad-hoc computing resources for performance, minimizing infrastructure hosting redundancy and complexity (https://aws.amazon.com/).

GeneLab provides a RESTful Application Programming Interface (API) to its full-text search engine (https://api.nasa.gov/api.html#genelab), which provides the same functionality available through the GLDS search interface. The API provides a choice of standardized web output formats, such as JavaScript Object Notation (JSON) or Hyper Text Markup Language (HTML), for the search results.

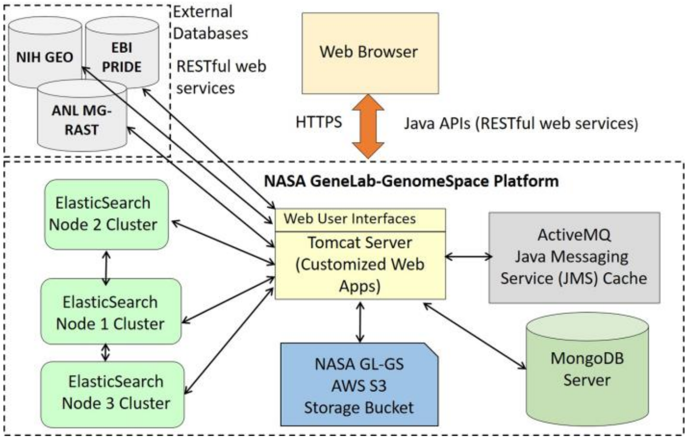

NIH GEO
EBI PRIDE
ANL MG-RAST
External Databases
RESTful web services
Web Browser
HTTPS
Java APIs (RESTful web services)
NASA GeneLab-GenomeSpace Platform
ElasticSearch Node 2 Cluster
ElasticSearch Node 1 Cluster
ElasticSearch Node 3 Cluster
Web User Interfaces
Tomcat Server (Customized Web Apps)
ActiveMQ Java Messaging Service (JMS) Cache
NASA GL-GS AWS S3 Storage Bucket
MongoDB Server

Figure 4. GLDS System Architectural Components (Customized GeneLab-GenomeSpace Platform). Graphical representation of technical coordination between external databases and web browser to the GeneLab-GenomeSpace platform utilizing the MongoDB server and NASA AWS storage.

4 Discussion

Since its launch in 2015 for initial Phase 1, GLDS has been providing users with an efficient data retrieval system equipped with federated and customizable full-text search capabilities, intuitive navigation, user interfaces using multiple tabs and organized detailed metadata, collaborative workspace and links to analysis tools for targeted space biology data analysis and collaborative result sharing. Although GeneLab is sponsored by NASA, both international and non-NASA funded investigations have been an integral part of the repository.

Flight samples archived at NASA Ames LSDA or obtained from the NASA Biospecimen Sharing Program (https://lsda.jsc.nasa.gov/Biospecimen) can also be processed by the GeneLab Sample Processing Laboratory (SPL). These samples are typically frozen archived spaceflight samples left-out from previous experiments and complementing already published set by principal investigators, or they are tissue of the same type and species reserved by GeneLab for data consistency and long-term longitudinal assessment. These samples are processed by wet-lab scientists in the SPL for extraction of DNA, RNA, and protein. The SPL team performs in-house sequencing (e.g. epigenomics, transcriptomics, metagenomics) or sends out extracts to core centers for the generation of other types of omics (e.g., metabolomics, proteomics). The SPL therefore plays a critical role in reducing noise and batch effects in the database by generating more consistent data as they apply a systematic standard operating procedure. By keeping up with state of the art technology, the SPL can also maintain and develop standardized processes, and recommend these processes for the entire Space Biology community, so that all datasets can be more easily combined and analyzed as a whole. This is very much in the spirit of larger multi-omics projects, such as the Cancer Genome Atlas where each type of omics had to be conducted by a single center (Tomczak, et al., 2015). GeneLab has integrated spike-in controls in their process, in collaboration with NIST (National Institute Standards and Technology), providing a systematic way to measure technical performance and data reproducibility (Munro, et al., 2014). The SPL thus solidifies Genelab's role in curating spaceflight omics data as well as archiving the cellular extracts from the expensive flight experiments.

The GeneLab platform has already enabled critical insights into the key factors affecting space flown organisms. In the past year, GeneLab scientists carried out a pan tissue (muscle, breast and liver) analysis highlighting the importance of the animal enclosure modules (AEM) used in rodent experiments on Bion and STS (Beheshti, et al., 2018). AEM enclosures were found to reduce metabolism, alter immune response and activate potential tumorigenic pathways as compared to the ground vivarium control animals. Another study focusing on meta-analysis of transcriptomic data from seven rodent data sets involving several tissues such as liver, kidney, adrenal gland, thymus, mammary gland, skin, and skeletal muscle (soleus, extensor digitorum longus, tibialis anterior, quadriceps, and gastrocnemius) helped identify key genes such as p53, TGFβ1, immune related genes, coordinating a global systemic response to microgravity (Beheshti, et al., 2018).

Such studies have highlighted the requirement for availability of detailed and complete metadata, accessibility to confounding stressors from spaceflight environment as well as increased sample size for stronger statistical reliability. GeneLab's continuous efforts will be directed towards employing evolved community standards for data submission, facilitating easier access to public spaceflight datasets, and targeted hypothesis-driven omics data analysis to strengthen space research around the globe.

5 Future developments

The GeneLab project is currently working on integrating Galaxy (Afgan, et al., 2016), an open source bioinformatics data analysis platform, with tools for transcriptomic, proteomic, epigenomics and metagenomic data analysis across all model systems. This will enable scientific researchers, as well as, bioinformatics enthusiasts with little programming background to be able to access spaceflight "omics" data, analyze and browse through higher order processed data and share with the broader scientific community. In an effort to develop higher community standards, GeneLab also organized the first Analysis Working Group (AWG) targeted for each model organism aiming towards establishment of optimized workflows for "Omics" data analysis to synchronize results and capitalize on the insights gained from the expensive and sparse datasets. More information on GeneLab AWG can be found at: https://genelab.nasa.gov/awg/charter.

With growing volume of data, GeneLab is also strategizing to develop internal tools to automatize the curation and review process. This includes developing submission algorithms guiding data submitters to provide specific information in order to validate the completeness of metadata and assess its relevance to our database. Concomitantly increased resources for computing and storage have been put in place to sustain such effort without any anticipated stop in the future.

Since the spaceflight community is relatively small and flight missions are expensive, GeneLab is working towards providing organized collective solutions such as processed data along with raw data files for each datasets. These results can be easily uploaded on the workspace for visualization with several types of graphical plots, such as principal component analysis (e.g., how experimental groups segregate), heat maps (e.g., extent of gene regulation), volcano plots (e.g., magnitude of expression for the top genes) and other pathway/network analysis tools. In addition to current experimental and sample processing metadata, GeneLab plans to include ISS environmental and Payload sensor detected environmental metadata, in a standardized fashion, for each flight dataset. Such information includes absorbed radiation (already being added), CO2 and temperature levels, routinely monitored in ISS. Inclusion of these additional features will help synchronize space associated biological data to be available to a wider community of space enthusiasts and engage multiple types of biological researchers towards novel hypothesis driven investigation.

Acknowledgements

The authors would like to acknowledge the NASA GeneLab Data Systems Team lead by Peter Tran and the Science team lead by Jonathan Galazka for their work on developing and staging the database and obtaining raw data from spaceflight samples.

Funding

Research funding was provided by the GeneLab Project at NASA Ames Research Center, through NASA's Space Biology Program in the Division of Space Life and Physical Sciences Research and Applications (SLPSRA). Any use of trade names is for descriptive purposes only and does not imply endorsement by the US Government.

References

Afgan, E., *et al.* The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic acids research* 2016;44(W1):W3-W10.

Allen, D.L., *et al.* Effects of spaceflight on murine skeletal muscle gene expression. *Journal of Applied Physiology* 2009;106(2):582-595.

Andarawewa, K.L., *et al.* Lack of radiation dose or quality dependence of epithelial-to-mesenchymal transition (EMT) mediated by transforming growth factor beta. *Int J Radiat Oncol Biol Phys* 2011;79(5):1523-1531.

Barcellos-Hoff, M.H., *et al.* Concepts and challenges in cancer risk prediction for the space radiation environment. *Life sciences in space research* 2015;6:92-103.

Barcellos-Hoff, M.H. and Ravani, S.A. Irradiated mammary gland stroma promotes the expression of tumorigenic potential by unirradiated epithelial cells. *Cancer Res* 2000;60(5):1254-1260.

Beheshti, A., *et al.* Global transcriptomic analysis suggests carbon dioxide as an environmental stressor in spaceflight: A systems biology GeneLab case study. *Scientific reports* 2018;8(1):4191.

Beheshti, A., *et al.* NASA GeneLab Project: Bridging Space Radiation Omics with Ground Studies. *Radiation research* 2018.

Beheshti, A., *et al.* A microRNA signature and TGF-β1 response were identified as the key master regulators for spaceflight response. *PloS one* 2018;13(7):e0199621-e0199621.

Blaber, E.A., *et al.* Microgravity induces pelvic bone loss through osteoclastic activity, osteocytic osteolysis, and osteoblastic cell cycle inhibition by CDKN1a/p21. *PloS one* 2013;8(4):e61372.

Chancellor, J.C., Scott, G.B. and Sutton, J.P. Space radiation: the number one risk to astronaut health beyond low earth orbit. *Life* 2014;4(3):491-510.

Gambara, G., *et al.* Gene expression profiling in slow-type calf soleus muscle of 30 days space-flown mice. *PloS one* 2017;12(1):e0169314.

Hei, T.K. Mitochondrial damage and radiation carcinogenesis. *Journal of radiation research* 2014;55(suppl_1):i17-i17.

Herranz, R., *et al.* Ground-based facilities for simulation of microgravity: organism-specific recommendations for their use, and recommended terminology. *Astrobiology* 2013;13(1):1-17.

Heuskin, A.C., *et al.* Simulating Space Radiation-Induced Breast Tumor Incidence Using Automata. *Radiat Res* 2016;186(1):27-38.

Jonscher, K.R., *et al.* Spaceflight activates lipotoxic pathways in mouse liver. *PloS one* 2016;11(4):e0152877.

Koike, Y., *et al.* Effects of HZE particle on the nigrostriatal dopaminergic system in a future Mars mission. *Acta astronautica* 2005;56(3):367-378.

Kwon, T., *et al.* Transcriptional response of Arabidopsis seedlings during spaceflight reveals peroxidase and cell wall remodeling genes associated with root hair development. *American Journal of Botany* 2015;102(1):21-35.

Maxwell, C.A., *et al.* Targeted and nontargeted effects of ionizing radiation that impact genomic instability. *Cancer Res* 2008;68(20):8304-8311.

Mayer-Wagner, S., *et al.* Effects of single and combined low frequency electromagnetic fields and simulated microgravity on gene expression of human mesenchymal stem cells during chondrogenesis. *Archives of medical science: AMS* 2018;14(3):608.

Missirian, V., *et al.* High atomic weight, high-energy radiation (HZE) induces transcriptional responses shared with conventional stresses in addition to a core "DSB" response specific to clastogenic treatments. *Frontiers in plant science* 2014;5:364.

Munro, S.A., *et al.* Assessing technical performance in differential gene expression experiments with external spike-in RNA control ratio mixtures. *Nature communications* 2014;5:5125.

Neumann, E. A life science Semantic Web: are we there yet? *Sci. STKE* 2005;2005(283):pe22-pe22.

Pecaut, M.J., *et al.* Is spaceflight-induced immune dysfunction linked to systemic changes in metabolism? *PloS one* 2017;12(5):e0174174.

Qu, K., *et al.* Integrative genomic analysis by interoperation of bioinformatics tools in GenomeSpace. *nature methods* 2016;13(3):245.

Rabin, B.M., *et al.* Effects of heavy particle irradiation and diet on object recognition memory in rats. *Advances in Space Research* 2009;43(8):1193-1199.

Sansone, S.-A., *et al.* Toward interoperable bioscience data. *Nature genetics* 2012;44(2):121.

Sridharan, D.M., *et al.* Understanding cancer development processes after HZE-particle exposure: roles of ROS, DNA damage repair and inflammation. *Radiat Res* 2015;183(1):1-26.

Tang, j., *et al.* Mathematical Modeling for DNA Repair, Carcinogenesis and Cancer Detection. In, *Genomic Instability and Cancer Metastasis*. Springer Science & Business Media; 2015. p. 75–93.

Tomczak, K., Czerwinska, P. and Wiznerowicz, M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp Oncol (Pozn)* 2015;19(1A):A68-77.

Weintraub, N.L., Jones, W.K. and Manka, D. Understanding radiation-induced vascular disease. In.: Journal of the American College of Cardiology; 2010.

Wilson, J., *et al.* Space flight alters bacterial gene expression and virulence and reveals a role for global regulator Hfq. *Proceedings of the National Academy of Sciences* 2007;104(41):16299-16304.

Worthington, J.J., *et al.* Regulation of TGFβ in the immune system: an emerging role for integrins and dendritic cells. *Immunobiology* 2012;217(12):1259-1265.